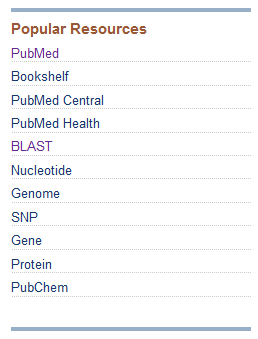
**Introduction To Bioinformatics:**

**Exercise 1:**

You will be given a nucleotide sequence found in real human DNA that is associated with a genetic disease when mutated. Your job is to compare the sequences you are given with the nucleotide sequence of most known genes, using the BLAST tool to search genetic databases.

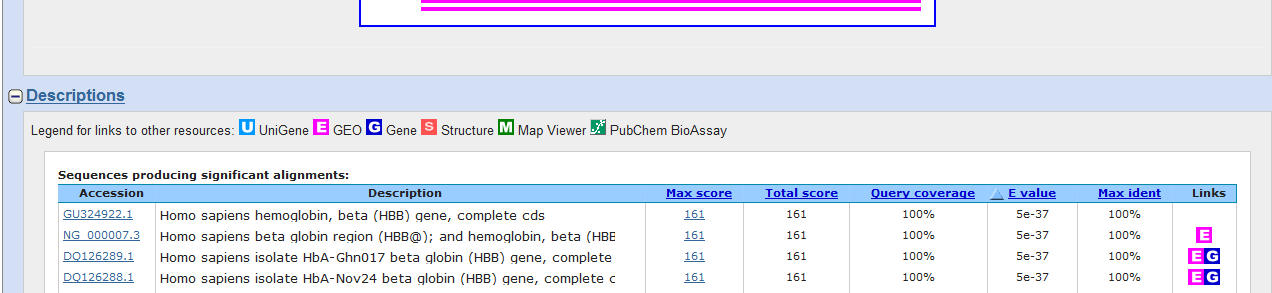
1. Go to the homepage for the NCBI ( [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) )
2. Click on the word "BLAST" located in the list of links on the right of the page.
3. Scroll down until you find the heading "**Nucleotide BLAST**" and click the link.



1. This time we will practice using a part of the gene sequence that codes for hemoglobin. When a mutation occurs in this gene a person can wind up with Sickle Cell Anemia. Cut and Paste the sequence below into the blast window.

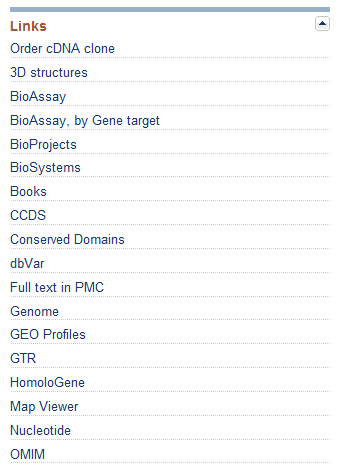
**GGG ATG AAT AAG GCA TAT GCA TCA GGG GCT GTT GCC AAT GTG CAT TAG CTG TTT GCA GCC TCA CCT TCT TTC ATG GAG TTT AAG ATA**



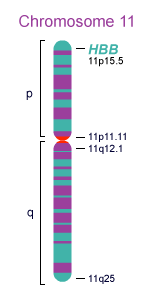
1. When you have finished entering your sequence, click the radio button for OTHERS and then click on the BLAST button.
2. You will then see a screen asking you to wait 10-20 seconds. Don’t click on anything just relax and wait patiently for the search to conclude.
3. After the search has ended, scroll down past the box with all of the pink bars in it and find the words **"Sequences Producing Significant Alignment".**
4. Listed in order are the closest matches with your DNA sequence. You should notice that the blast search will return results for all genomes currently mapped (Several prokaryotes, humans, rats, chimpanzees, cows, pigs, chickens and puffer fish to name a few)
5. Please take a moment to be awed by the similarity found in the DNA code despite the outward physical diversity of organisms.
6.  Click on the blue reference number (right side) attached to the second match. This will tell you the name of the gene and its abbreviation if available).
7. Scroll down to “Related Info” on the right side and click “Gene”

|  |
| --- |
|  |

1. This is the Gene Summary. It tells you that this sequence is officially recognized. HBB is the official Symbol for this gene. If you read further it will tell you that it is found in humans and that a mutated version of it causes Sickle Cell Anemia.
2. On the right you will see a link to PHENOTYPES. Click the link and look for information about Hb SS Disease under summary from GeneReviews. This will open more information about sickle cell.



1. Return to the Gene Summary page. Next click on the link to the “Map Viewer” in the right column under the “Related Sites” heading.



1. Map Viewer page. Across the top of the page, you will notice the numbers 1-22 XY. These numbers represent the chromosomes found in humans.
2. While we are here lets take a moment to go over chromosomal nomenclature

**11p15.5–** This means that the HBB gene is found on chromosome 11 on the short (p) arm in region 15.5. the fancy word for this is Locus which is science talk for location.



1. Your next step will be to click on the OMIM link in the information bar. This will take you to a page with lots of information about your gene and what it does.

**Exercise 2:** Query a public database

*Examination of a GenBank entry*

* Go to the [Genbank website](http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide)
* Search for “DMD AND Homo sapiens”
* Select the RefSeq tab (left side)
* Select transcript variant Dp427m, (accession code NM\_004006)
* Explore the information, e.g. gene, coding sequence (CDS) , PubMed
* Find the variation in region 3713. (scroll way down) **What effect has this variant?**

*Retrieve a sequence from the database and store it locally*

The "fasta" format is a format for sequences, which can be read by many programs.

* Search in the Genbank database for 'human mRNA HBA2' (NM\_000517) Tip: you can search on accession code
* Set Display (top left of the page) to Fasta
* Save as a text file, save as NM\_000517.txt (This is only to show how you can save information from GenBank, we do not need this sequence anymore)

**What are the characteristics of the FASTA format?**

**Exercise 3:** Compare two sequences to identify mutations

* Read more about the Sickle cell anemia gene [here](http://ghr.nlm.nih.gov/condition/sickle-cell-disease) and [here.](http://ghr.nlm.nih.gov/gene/HBB)
* **Which mutation causes Sickle cell disease?**

The bl2seq program (BLAST) is used to align two sequences against each other instead of against the entire database. The sequence can be pasted into the boxes or the accession codes (unique identifier for a sequence) can be entered as input.

* Go to the [BLAST website](http://www.ncbi.nlm.nih.gov/BLAST/). Determine the position of the HBS mutation with the program "Align two sequences (bl2seq)" (In the "Specialized Blast" box)
* Type "NM\_000518" in the box for sequence1, this is the mRNA sequence of normal HBB.
* Type "M25113" (mRNA of HBS) in the second box.
* Click on "Blast" and examine the output.
* At the top of the page you can alter the 'formatting options'. Tick the "CDS feature" box and click on "Reformat" to see the protein translations of the sequences.

The sequences are retrieved from the nucleotide database (GenBank) and are placed above each other. This is called an alignment. Matching nucleotides are indicated by a pipe-sign: |. Mismatching nucleotides will not have a pipe-sign. Insertions/deletions are annotated as a dash (-) in one of the sequences. Below the alignment the protein sequence is placed when the annotation of one of the sequences is known. In this case the identifiers were given as input, so the bl2seq program could retrieve this information from GenBank.

* **Identify the mutation that causes the change in amino acid.**

**Exercise 4:** Similarity search with BLAST

When considering a new DNA sequence, the first question will probably be: "*What do I have in my sequence?*" which relates to the question "*Is my sequence similar to previously known sequences?*" To find out, you may compare your sequence with the public databases, using one or more computer programs that have been written for searching the databases for similarities to a given query sequence. The fastest way to search for similar sequences in the databases is **BLAST** (Basic Local Alignment Search Tool). Blast is a set of similarity search programs designed to explore all of the available [sequence databases](http://www.ncbi.nlm.nih.gov/BLAST/blast_databases.shtml) regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. Blast has several variants:

|  |  |  |
| --- | --- | --- |
| [**Sequence type**](http://www.bioinformaticslaboratory.nl/twiki/bin/view/BioLab/DnaTechnology?sortcol=0;table=2;up=0#sorted_table) | [**nucleotide database**](http://www.bioinformaticslaboratory.nl/twiki/bin/view/BioLab/DnaTechnology?sortcol=1;table=2;up=0#sorted_table) | [**amino acid database**](http://www.bioinformaticslaboratory.nl/twiki/bin/view/BioLab/DnaTechnology?sortcol=2;table=2;up=0#sorted_table) |
| **nucleotide query** | blastn/tblastx | blastx |
| **amino acid query** | tblastn | blastp |

* blastn is good for finding nucleotide sequences similar to yours.
* blastp is good for finding amino acid sequences similar to yours.
* blastx is good for finding amino acid sequences similar to any translation of your nucleotide sequence - e.g. if you could not recognize an ORF.
* tblastn is good for finding nucleotide sequences that can be translated into something similar to your amino acid sequence - e.g. unannotated pseudogenes.
* tblastx is good for keeping computers busy (or for very specific applications).

The BLAST 'Search' box accepts a number of different types of input and automatically determines the format. Accepted input types are: FASTA file, bare sequence or database identifiers (e.g. genbank i.d. M18533, Swissprot i.d. P11532)

* Open [BLAST](http://www.ncbi.nlm.nih.gov/BLAST) and go to *blastp (protein Blast)*
* Open the file “codes for exercises 1, 4 and 5” (on Schoology). Copy and paste the exercise 4 sequence into the sequence input field on the *blastp* page, leave the other options unchanged and press BLAST. (This sequence is translated from an mRNA sequence)
* **Which sequence in the database is most similar to the one we have submitted? Are the sequences identical?**

A Graphical overview of the database sequences aligned to the query sequence is shown. The score of each alignment is indicated by one of five different colors, with the most similar hits uppermost and appearing in red. Pink, green, blue and black bars follow, representing proteins in decreasing order of similarity. Hatched areas would indicate a gap in similarity i.e., two or more distinct regions of similarity were found within the same sequence hit. Detached bars on the same line correspond to unrelated hits. Mousing over a hit sequence causes the definition and score to be shown in the window at the top, clicking on a hit sequence takes the user to the associated alignments.

* On the Blast result page click on Taxonomy reports and look at the hits for different organisms
* Go back to the Blast result page. You can access different databases by clicking on the accession code. Explore some links.
* Find the original protein, transcript variant Dp116 (NP\_004005). From which gene does this protein originate? **Find the RefSeq accession code for the corresponding mRNA sequence.**
* In the database record for NM\_004014, explore the links in the menu on the right. You can do further analysis with this sequence and link out to other databases.
* Try to find PCR primers that could be used for Dp116. There’s a link “Pick Primers” on the right side. That will take you to the Primer picking page. Let’s not worry about customizing, just scroll down to “Get Primers” and click. (This takes awhile. Select the “open in new tab” option so the wait happens in a different tab.

**Bioinformatics Exercise 5:** Pairwise sequence alignment

Pairwise sequence alignment methods are concerned, in contrast to BLAST, with finding the best-matching piecewise local or global alignments of protein and DNA sequences.

*Local alignment (Smith-Waterman algorithm)*

Local alignment methods find related regions within sequences - in other words they can consist of a subset of the characters within each sequence (e.g. positions 20-40 of sequence A might align with positions 50-70 of sequence B).This is a more flexible technique than global alignment and has the advantage that related regions which appear in a different order in the two proteins (which is known as domain shuffling) can be identified as being related.

*Global alignment (Needleman-Wunsch algorithm)*

A global alignment between two sequences is an alignment in which all of the characters in both sequences participate in the alignment. Global alignments are useful mostly for finding closely-related sequences.

**EMBOSS-Align**

The EMBOSS-Align tool contains two programs each using a different algorithm:

* When you want an alignment that covers the whole length of both sequences, use the needle program
* When you are trying to find the best region of similarity between two sequences, use the water program

Open [EMBOSS](https://www.ebi.ac.uk/Tools/emboss/) and then select the “protein” option under “Water” (the local alignment)

* Compare the protein sequences for the Human dystrophin protein with that of the chicken dystrophin protein. You can also try this again with the mouse and the fruit fly sequences.
* Repeat this process but use “Water” (the global alignment)
* **Which alignment is better?**

Note that matching amino acids are connected with a "|" symbol. Mismatches would be connected with a space. A gap would be represented with a "-" symbol. Similar amino acids (e.g. threonine vs methionine) are connected via a "." symbol. Thus a sequence alignment can be represented in the format...

DMFCNTEGQGIAMM

| |||| ..

TMG--NEGQGSETT

**Exercise 6:** GeneCards database for finding information about a gene

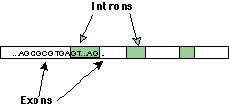
In agreement with evolutionary principles, scientific research to date has shown that all genes share common elements. For many genetic elements, it has been possible to construct consensus sequences, those sequences best representing the norm for a given class of organisms (e.g, bacteria, eukaroytes). Common genetic elements include promoters, enhancers, polyadenylation signal sequences and protein binding sites.

A mRNA can be divided into three parts: a 5′ untranslated region (5′ UTR), the polypeptide coding region, sometimes called the open reading frame (ORF), and the 3′ translated region (3′ UTR).

The first codon in a messenger RNA sequence is almost always AUG. While this reduces the number of candidate codons, the reading frame of the sequence must also be taken into consideration.

There are six reading frames possible for a given DNA sequence, three on each strand, that must be considered, unless further information is available. Since genes are transcribed away from their promoters, the definitive location of this element can reduce the number of possible frames to three. The location of the appropriate start codon will include a frame in which there are no apparent abrupt stop codons. Incorrect reading frames usually predict relatively short peptide sequences. Therefore, it might seem deceptively simple to ascertain the correct frame. In bacteria, such is frequently the case. However, eukaryotes add a new obstacle to this process, introns.

Intron/exon splice sites can be predicted on the basis of their common features. Most introns begin with the nucleotides GT and end with the nucleotides AG.



1. Have a look at the organization of introns/exons on the [DMD homepage](http://www.dmd.nl/seqs/murefDMD.html)

2. GeneCards is a database of human genes, their products and their involvement in diseases and there are several ways to search:

* Go to the [GeneCards website](http://www.genecards.org/) .
* Type DMD into the search box.
* Click on “Disorders” under the search box.

This card shows all information available on the gene, with links to corresponding databases, e.g.

3. Search OMIM <http://www.ncbi.nlm.nih.gov/omim> for DMD

**Exercise 7:** Gene finding

The **ORF Finder** (Open Reading Frame Finder) is a graphical analysis tool that finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.

* Open [ORF finder](http://www.ncbi.nlm.nih.gov/gorf/gorf.html)
* Enter the accession number "NM\_004006" (Dystrophin transcript variant Dp427m) into the appropriate box.

You will see this full cDNA sequence translated in all 6 reading frames with the results shown in diagrammatic format. Remember that DNA is double stranded so there are 3 reading frames for the top strand and 3 more reading frames for the bottom strand. All the ORFs are indicated by the colored boxes with the rest of the cDNA left uncolored. The ORF Finder is a graphical-analysis tool which allows you to find all open reading frames in a given sequence

|  |  |
| --- | --- |
|  | Open Reading Frame from the list |
|  | Selected Open Reading Frame |
|  | Accepted Open Reading Frame |

What you are generally looking for is the largest ORF. The software has already done this for you by ranking the ORFs from largest to smallest. It also tells you which frame it came from, for example +3 means it is the third frame (3) on the top strand (+).

* Click on the colored box next to the +2 from the list on the largest ORF. You will see the color of the ORF box change and the deduced amino acid sequence.
* "Accept" the proposed ORF and View Genbank entry and find the Gene name
* Copy/paste the [NM\_004006\_point.txt](http://amc-app1.amc.sara.nl/twikidata/pub/Education/DNA_technology/NM_004006_point.txt) sequence in ORF finder and compare the output with the previous one

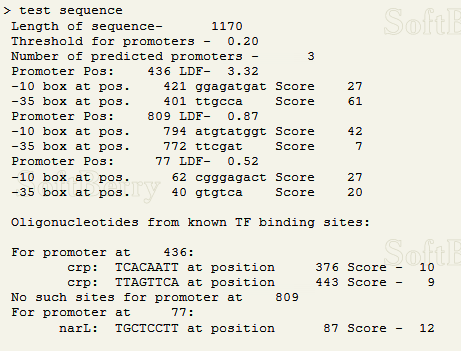
**What is wrong with this transcript?**

**Exercise 8:** Promotor prediction.

In this exercise we will use the BPROM program to identify the promotor region of the hep2 gene of *Heliobacillus mobilis.*

1. Retrieve the sequence containing the hep2 gene from [GenBank](http://www.ncbi.nlm.nih.gov/genbank/) (accession code EU052681).
2. Click on “fasta” )upper left of page) to convert the sequence to a fasta file.
3. The BPROM algorithm predicts potential transcription start positions of bacterial genes regulated by sigma70 promoters (major E.coli promoter class). The σ70 (RpoD) factor is the "housekeeping" sigma factor that transcribes most genes in growing cells. A linear discriminant function (LDF) is used that predict these transcription start sites, and which combines characteristics describing functional motifs and oligonucleotide composition of these sites.

Go to the [BPROM](http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb) website and use your sequence as input for promotor prediction. The result is shown in the figure below.



**Figure 6.** Output of the BPROM program for the hep2 gene of *Heliobacillus mobilis*

**Exercise 9: Use BLAST to find out about (at least) 3 of these genes:**

**Gene Sequence 1**

ATG GCG ACC CTG GAA AAA GCT GAT GAA GGC CTT CGA GTC CCT CAA GTC CTT CCA GCA GCA GCA GCA GCA GCA GCA GCA GCA GCA GCA GCA GCA GCA GCA GCA GC

**Gene Sequence 2**

ATG GCG GGT CTG ACG GCG GCG GCC CCG CGG CCC GGA GTC CTC CTG CTC CTG CTG TCC ATC CTC CAC CCC TCT CGG CCT GGA GGG GTC CCT GGG GCC ATT CCT GGT GGA GTT CCT GGA GGA GTC TT

**Gene Sequence 3**

ATG CTC ACA TTC ATG GCC TCT GAC AGC GAG GAA GAA GTG TGT GAT GAG CGG ACG TCC CTA ATG TCG GCC GAG AGC CCC AGC CCG CGC TCC TGC CAG GAG GGC AGG CAG GGC CCA GAG GAT GGA G

**Gene Sequence 4**

ATG TTT TAT ACA GGT GTA GCC TGT AAG AGA TGA AGC CTG GTA TTT ATA GAA ATT GAC TTA TTT TAT TCT CAT ATT TAC ATG TGC ATA ATT TTC CAT ATG CCA GAA AAG TTG AAT AGT ATC AGA TTC CAA ATC T

**Gene Sequence 5**

ATG CGT CGA GGG CGT CTG CTG GAG ATC GCC CTG GGA TTT ACC GTG CTT TTA GCG TCC TAC ACG AGC CAT GGG GCG GAC GCC AAT TTG GAG GCT GGG AAC GTG AAG GAA ACC AGA GCC AGT CGG GCC

**Gene Sequence 6**

ATG CCG CCC AAA ACC CCC CGA AAA ACG GCC GCC ACC GCC GCC GCT GCC GCC GCG GAA CCC GGC ACC GCC GCC GCC GCC CCC TCC TGA GGG ACC CAG AGC AGG ACA GCG GCC CGG AGG AC

Type your answers directly into the tables below

**#1**

|  |  |
| --- | --- |
| **Gene Sequence Number** |  |
| **Abbreviation or name of gene** |  |
| **Genetic disease associated with gene** |  |
| **Chromosome number (location of gene)** |  |
| **Describe symptoms of disease** |  |

**#2**

|  |  |
| --- | --- |
| **Gene Sequence Number** |  |
| **Abbreviation or name of gene** |  |
| **Genetic disease associated with gene** |  |
| **Chromosome number (location of gene)** |  |
| **Describe symptoms of disease** |  |

**#3**

|  |  |
| --- | --- |
| **Gene Sequence Number** |  |
| **Abbreviation or name of gene** |  |
| **Genetic disease associated with gene** |  |
| **Chromosome number (location of gene)** |  |
| **Describe symptoms of disease** |  |

**Analysis**

1. In a BLAST search how do you know which sequence of DNA was closest to your unknown sequence?

2. If you had more nucleotides in your sequence to enter into BLAST, (say 1000 instead of 100), do you think you would get more specific or less specific matches? Why?

3. Why do scientists all over the world check BLAST when they sequence a new region of DNA?

4. Why would a biologist studying human genetic disorders want to see related genetic entries for mice, cats, dogs etc?

5. How does the information in the NCBI help to illustrate the interrelation of all organisms? Explain.

6. What do you think the potential consequences are for insurance companies knowing DNA sequences of individuals? Should our genetic information be shared with companies, employers, future spouses, doctors? Why or Why not?

**Exercise 10: Use Blast, Gene Card and OMIM to search for genes of your choosing.**

You might try searching for genes such as myosin, actin, SRY, GFP, ATP synthase, Nitrate sensor, or Na+/K+ ATPase or anything you want to.

**Exercise 11: Using Bioinformatics to study phylogenetic relationships;**

Are polar bears closely related to American brown bears? Do you think the giant panda is more closely related to the American brown bear or the polar bear? Is the giant panda a bear at all? What about the red panda – is it a giant panda relative? What does the evolutionary history of bears look like? In this activity you will use real gene sequences to figure out which animals are more closely related. You will create a family tree for bears and help to solve the mystery of bear history!

**TABLE 1: Accession numbers and scientific names different species**

|  |  |  |
| --- | --- | --- |
| **Organism** | **Scientific Name** | **Accession Number** |
| American black bear |  | Y08520.1 |
| American brown bear |  | L21889.1 |
| Spectacled bear |  | L21883.1 |
| Asiatic bear |  | L21890.1 |
| Polar bear |  | L22164.1 |
| Giant panda |  | Y08521.1 |
| Red Panda |  | S80939.1 |
| Northern raccoon |  | U78345.1 |
| Leopard seal |  | AY377288.1 |

1. Go to [**www.ncbi.nlm.nih.gov**](http://www.ncbi.nlm.nih.gov). Click the arrow beside “all databases” and select **Nucleotide**. Enter the **accession number** for the American Black bear (see Table 1 for number) in the search space.
2. You will see lots of information. Near the top you will see the definition listed as: *U. americanus* mitochondrial 12S rRNA gene. *U. americanus* is the scientific name for the American black bear. The gene is part of the mitochondrial DNA and makes up part of each ribosome. Write the scientific name for the American black bear in Table 1.
3. Click **FASTA** at the top of the page. You will then see the sequence of DNA that makes up this 12S rRNA gene.
4. On the right, towards the top, is a drop-down arrow next to “Send”. Click the arrow and choose **Clipboard** as your destination. Then click “**Add to Clipboard**.”
5. Continue by searching each animal’s accession number, clicking on the FASTA, and sending the sequences to the clipboard.
6. After you have added all 9 FASTA sequences to the clipboard, click on the “**9 items**” next to clipboard to open your clipboard.
7. **Select each animal** by checking the box.
8. After all nine have been selected, click the “**Send to**” in the upper right hand corner.
9. Chose “**File**” and under Format select “**FASTA**”.
10. Chose “**Organism Name**” under “**Sort By**”.
11. Click “**Create File**”
12. Your file will be created with the following name: **sequence (1).fasta**, which will be automatically downloaded. You will be able to find this file in your download files.

You will now take these sequences for the 12S rRNA gene and align them using an online website. The website uses algorisms (math) to find areas of similarity across the 9 organisms. This requires significant computation so this is always done via computers.

1. Now you will go to [**http://www.genome.jp/tools/clustalw**](http://www.genome.jp/tools/clustalw).
2. Chose “**CLUSTAL**” for the output format.
3. Check “**SLOW/ACCURATE**” for the pairwise alignment.
4. Check “**DNA**” for Enter your sequences.
5. You will upload the sequence file you just made by choosing “**Choose File**”. Navigate to find the file you downloaded called **sequence (1).fasta**.
6. Click “**Execute Multiple Alignment**”.
7. The program will align the sequences to see the results.

***Interpreting your results***

1. At the top you will see the 9 sequences with their accession numbers and the number of base pairs (bp) for the sequence. The program numbered your organisms 1 – 9. In the table below, write the common name beside each number. You must look back at the accession numbers to match the right name to the correct number.

|  |  |
| --- | --- |
| **Organism Number** | **Organism Name** |
| **1** |  |
| **2** |  |
| **3** |  |
| **4** |  |
| **5** |  |
| **6** |  |
| **7** |  |
| **8** |  |
| **9** |  |

1. Now scroll down slightly and you will see alignment scores like below:

Sequences (1:2) Aligned. Score: 75.

1. This is just an example and will not match your data! This is the score between organism 1 and organism 2. The **higher** the **alignment score** between two organisms, the **more similar** their **DNA** is for this gene.
2. Scroll down farther and look at the sequence alignments; you will see nucleotide sequences (the A’s, G’s, T’s, and C’s). The symbol “–” indicates that there is no information for this location in the sequence. In other words the base (letter) is not known. Because some sequences are longer than others there may be many sections where information is missing. The symbol “**\***”indicates a location where all nine organisms have the exact same nucleotide, meaning the base is conserved. If a nucleotide is conserved then that same nucleotide should be in the common ancestor.

***QUESTIONS ABOUT THE ALIGNMENT***

1. Which two species have the highest alignment scores? You will need to check the accession numbers so you know which organism is which.
2. Which two species have the lowest alignment scores?
3. Which group has the highest score with the red panda?
4. Which group has the second highest score with the red panda?
5. Which group has the lowest score with the red panda?
6. What does it mean if two groups have a high alignment score?
7. Now you will use this website to build a phylogenic tree. This is like a family tree based on genetics and evolution. The website takes the data that you have looked at and draws a tree, which is more visually appealing.
8. At the top of the page (above the alignment data), select “Rooted phylogenetic tree with branch length (UPGMA)” from the “Select tree menu” drop-down menu.
9. Click on “**Exec**”.
10. You will see the phylogenic tree. What does this tell you about the relationships among bear species?

**Exercise 12 “Extra Credit”: Can you figure out what species these proteins belong to? Can you create a phylogenetic tree? Can you find the DNA sequences for these proteins?**

Species One: Beta-hemoglobin Amino Acid Sequence

1. mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk

61 vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfrllgnv lvcvlahhfg

121 keftppvqaa yqkvvagvan alahkyh

Species Two: Beta-hemoglobin Amino Acid Sequence

1 mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk

61 vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfrllgnv lvcvlahhfg

121 keftppvqaa yqkvvagvan alahkyh

Species Three: Beta-hemoglobin Amino Acid Sequence

1 mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk

61 vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfkllgnv lvcvlahhfg

121 keftppvqaa yqkvvagvan alahkyh

Species Four: Beta-hemoglobin Amino Acid Sequence

1 vhltpeeksa vtalwgkvnv devggealgr llvvypwtqr ffesfgdlst pdavmgnpkv

61 kahgkkvlga fsdglahldn lkgtfaqlse lhcdklhvdp enfrllgnvl vcvlahhfgk

121 eftpqvqaay qkvvagvana lahkyh

Species Five: Beta-hemoglobin Amino Acid Sequence

1 vhltpeekna vttlwgkvnv devggealgr llvvypwtqr ffdsfgdlss paavmgnpkv

61 kahgkkvlga fsdglnhldn lkgtfaqlse lhcdklhvdp enfkllgnvl vcvlahhfgk

121 eftpqvqaay qkvvagvana lahkyh

Species Six: Beta-hemoglobin Amino Acid Sequence

1 vhltgeeksa vttlwgkvnv eevggealgr llvvypwtqr ffesfgdlss pdavmgnpkv

61 kahgkkvlga fsdglahldn lkgtfaqlse lhcnklhvdp enfrllgnvl vcvlahhfgk

121 eftpqvqaay qkvvagvana lahkyh

Species Seven: Beta-hemoglobin Amino Acid Sequence

1 vhltgeeksa vttlwgkvnv eevggealgr llvvypwtqr ffesfgdlss pdavmgnpkv

61 kahgkkvlga fsdglahldn lkgtfaqlse lhcnklhvdp enfrllgnvl vcvlahhfgk

121 eftpqvqaay qkvvagvana lahkyh